EXMARaLDA und Datenbank ‚Mehrsprachigkeit' – Konzepte und praktische Erfahrungen*

Thomas Schmidt

SFB 538 ‚Mehrsprachigkeit', Universität Hamburg

This paper presents some concepts and principles used in the development of a database of multilingual spoken discourse at the University of Hamburg. The emphasis of the first part is on general considerations for the handling of heterogeneous data sets: After showing that diversity in transcription data is partly conceptually and partly technologically motivated, it is argued that the processing of transcription corpora should be approached via a three-level architecture which separates form (application) and content (data) on the one hand, and logical and physical data structures on the other hand. Such an architecture does not only pave the way for modern text-technological approaches to linguistic data processing, it can also help to decide where and how a standardization in the work with heterogeneous data is possible and desirable and where it would run counter to the needs of the research community. It is further argued that, in order to ensure user acceptance, new solutions developed in this approach must take care not to abandon established concepts too quickly.

The focus of the second part is on some practical experiences with users and technologies gained in the four years' project work. Concerning the practical development work, the value of open standards like XML and Unicode is emphasized and some limitations of the "platform-independent" JAVA technology are indicated. With respect to users of the EXMARaLDA system, a predominantly conservative attitude towards technological innovations in transcription corpus work can be stated: individual users tend to stick to known functionalities and are reluctant to adopt themselves to the new possibilities. Furthermore, an active commitment to cooperative corpus work still seems to be the exception rather than the rule.

It is concluded that technological innovations can contribute their share to a progress in the work with heterogeneous linguistic data, but that they will have to be supplemented, in the long run, with an adequate methodological reflection and the creation of an appropriate infrastructure.

---

* Ich danke den Teilnehmern des Workshops für die fruchtbaren Diskussionen.

Interdisciplinary Studies on Information Structure 02 (2005): 21–42
Dipper, S., M. Götze and M. Stede (eds.):
Heterogeneity in Focus: Creating and Using Linguistic Databases

1 Einleitung

In diesem Aufsatz geht es um die Datenbank ‚Mehrsprachigkeit' und das System EXMARaLDA, die am SFB 538 ‚Mehrsprachigkeit' der Universität Hamburg entwickelt werden. Da deren konzeptuelle und technische Details bereits an anderer Stelle ausführlich dargestellt worden sind (z.B. Schmidt 2004), soll der Schwerpunkt hier einerseits auf solchen Aspekten liegen, die – gemäß dem Thema des Workshops – mit allgemeineren Fragen zum Umgang mit computerverwertbaren, heterogenen linguistischen Datenbeständen zu tun haben. Andererseits soll versucht werden, aus den praktischen Erfahrungen der nunmehr vierjährigen Projektarbeit einige Erkenntnisse abzuleiten, die über den konkreten Projektzusammenhang hinaus für die weitere Arbeit auf diesem Gebiet interessant sein könnten.

2 Daten am SFB ‚Mehrsprachigkeit'

2.1 Überblick

Der Sonderforschungsbereich 538 „Mehrsprachigkeit" vereinigt in seinen vierzehn Teilprojekten eine Vielzahl von Forschern, die sich unter verschiedenen Herangehensweisen dem Thema der Mehrsprachigkeit widmen. In der derzeit laufenden zweiten Förderungsphase (2002-2005) ist der SFB in drei thematische Teilbereiche – „Erwerb der Mehrsprachigkeit", „Mehrsprachige Kommunikation" und „Historische Aspekte der Mehrsprachigkeit" – gegliedert. In ausnahmslos allen Projekten dieser Teilbereiche wird auf empirischer Basis gearbeitet, d.h. Ausgangspunkt der linguistischen Analysen bilden jeweils mehrsprachige Korpora geschriebener oder transkribierter gesprochener Sprache. Die fol-

genden Ausführungen beziehen sich auf die Korpora derjenigen Projekte, die mit gesprochener Sprache arbeiten, genauer:

Projekt	Sprachen	Arbeitsgebiet, theor. Hintergr.	Datentypen (Transk.-System)
K1: Japanische und deutsche Expertendiskurse	Japanisch Deutsch	Diskursanalyse Funktionale Pragmatik	Vortrags- und Planungsdiskurse (HIAT / syncWriter)
K2: Dolmetschen im Krankenhaus	Portugiesisch Türkisch Deutsch	Diskursanalyse Funktionale Pragmatik	Gedolmetschte Arzt-Patienten-Gespräche (HIAT / syncWriter)
K5: Semikommunikation und rezeptive Mehrsprachigkeit im heutigen Skandinavien	Dänisch Schwedisch Norwegisch	Diskursanalyse Funktionale Pragmatik	Radiosendungen, Interviews, Gruppen- und Unterrichtsgespräche (HIAT / HIAT-DOS)
E2: Simultaner und sukzessiver Erwerb von Mehrsprachigkeit	Französisch Portugiesisch Baskisch Spanisch Deutsch	Syntax Generative Grammatik	Spracherwerbsdaten (Interviewer-Kind-Interaktion) (LAPSUS)
E3: Prosodische Beschränkungen zur phonologischen und morphologischen Entwicklung im bilingualen Erstspracherwerb	Spanisch Deutsch	Phonologie Optimalitätstheorie	Spracherwerbsdaten (Interviewer-Kind-Interaktion) (IPA / WordBase)
E4: Spezifische Sprachentwicklungsstörung und früher L2-Erwerb	Deutsch Türkisch	Syntax Generative Grammatik	Spracherwerbsdaten (Interviewer-Kind-Interaktion) (DIGITRAIN / DACODA)
E5: Sprachliche Konnektivität bei bilingual türkisch-deutsch aufwachsenden Kindern	Türkisch Deutsch	Diskursanalyse Funktionale Pragmatik	Spracherwerbsdaten (Evokative Feldexperimente) (HIAT / syncWriter)

2.2 Heterogenität von Transkriptionsdaten

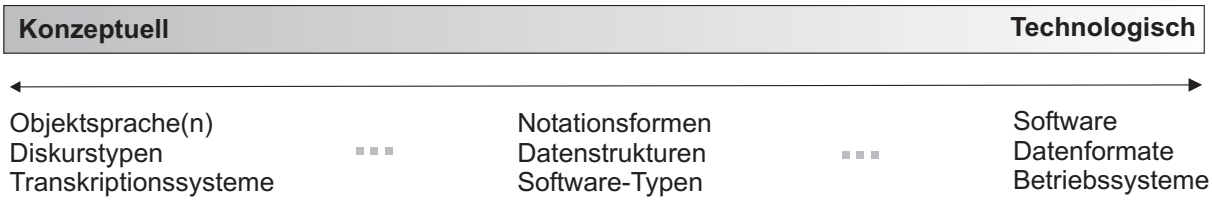
Die obige Tabelle deutet bereits an, dass hinsichtlich der Transkriptionsdaten am SFB eine große Heterogenität besteht. Diese resultiert teils aus *konzeptuellen* Motiven, also aus Unterschieden, die eher (gegenstands-)theoretisch begründet sind, teils aus *technologischen* Motiven, also aus Unterschieden, die eher mit der konkreten technischen Umgebung der praktischen Korpusarbeit zu tun haben.

Ein Beispiel für eine überwiegend konzeptuell bedingte Heterogenität findet sich in den *Transkriptionssystemen*, nach deren Vorgaben Aufnahmen gesprochener Sprache in digitale „Verschriftlichungen" überführt werden. Z.B. benötigen Projekte, die an diskursanalytischen Fragestellungen interessiert sind, für ihre Untersuchungen eine möglichst präzise und nachvollziehbare Repräsen-

tation des zeitlichen Gesprächsablaufs. Sie wählen daher ein Transkriptionssystem, das angemessene Mittel zum Abbilden simultaner Gesprächsbeiträge und non-verbaler Kommunikation beinhaltet – für die diskursanalytisch arbeitenden SFB-Projekte ist dies das mit der Partiturnotation arbeitende System HIAT (Rehbein et al. 2004). Für Projekte, deren Hauptaugenmerk auf phonologischen Fragestellungen liegt, ist hingegen vor allem eine möglichst exakte Repräsentation der lautlichen Gestalt isolierter Redebeiträge wichtig, während der genaue zeitliche Gesprächsablauf in der Transkription modellhaft ausgeblendet werden kann. Das betreffende Projekt E3 arbeitet daher mit einfachen Listen von IPA-transkribierten Äußerungen.

Ausschließlich technologisch bedingt ist hingegen ein Unterschied in der verwendeten *Transkriptionssoftware*: obwohl die Projekte K1, K2 und E5 einerseits und das Projekt K5 andererseits gleiche Arbeitsgebiete und theoretische Hintergründe aufweisen und folglich auch das gleiche Transkriptionssystem (HIAT) verwenden, bewerkstelligten sie die Transkriptionsarbeit ursprünglich mit unterschiedlicher Software (syncWriter bzw. HIAT-DOS); und dieser Unterschied lag ausschließlich in der Rechnerausstattung, genauer: in den jeweils bevorzugten Betriebssystemen (MAC OS bzw. Windows), der Projekte begründet – eine betriebssystemübergreifend einsetzbare Software zum Transkribieren in Partitur-Notation existierte nicht.

Wie die folgende Abbildung illustriert, ordnen sich weitere Aspekte der Heterogenität von Transkriptionsdaten auf einem Spektrum zwischen rein konzeptuell und rein technologisch motivierten Unterschieden ein.



Weiterhin sind diese diversen Aspekte der Heterogenität nicht unabhängig voneinander: Beispielsweise zieht die (rein konzeptuell motivierte) Wahl eines bestimmten Transkriptionssystems oft unweigerlich die (teils konzeptuell, teils technologisch motivierte) Wahl einer bestimmten Notationsform nach sich, die wiederum eine Software erforderlich machen mag, die nicht plattformübergreifend implementiert ist und somit zwangsläufig auch die (eigentlich rein technologisch zu motivierende) Wahl eines Betriebssystems vorbestimmt.

3 Prinzipien und Systemarchitektur der Datenbank ‚Mehrsprachigkeit'

Die Heterogenität der Transkriptionsdaten am SFB erschwert deren Austausch zwischen einzelnen Projekten und stellt somit ein Hindernis für die kooperative Forschung dar: es ist meist nicht ohne Weiteres möglich, die Projektdaten eines Projektes in den Arbeitsumgebungen eines anderen Projektes anzusehen oder auszuwerten, geschweige denn verschiedene Projektkorpora zu vereinen oder mit anderen als den ursprünglich vorgesehenen Werkzeugen zu bearbeiten. Darüber hinaus verhindert die Vielfalt der Formate eine einheitliche und effektive Archivierung der Daten und birgt so die Gefahr, dass aufwändig erstellte Korpora auf lange Sicht unbrauchbar werden.

Ziel des SFB-Projekts ‚Datenbank Mehrsprachigkeit' ist daher die Konzeption und Implementierung einer Plattform für die Erstellung und Auswertung von Korpora gesprochener Sprache, die die älteren projektspezifischen Lösungen ablösen und eine flexible Verarbeitbarkeit, Austauschbarkeit und Archivierbarkeit von Transkriptionsdaten gewährleisten soll.

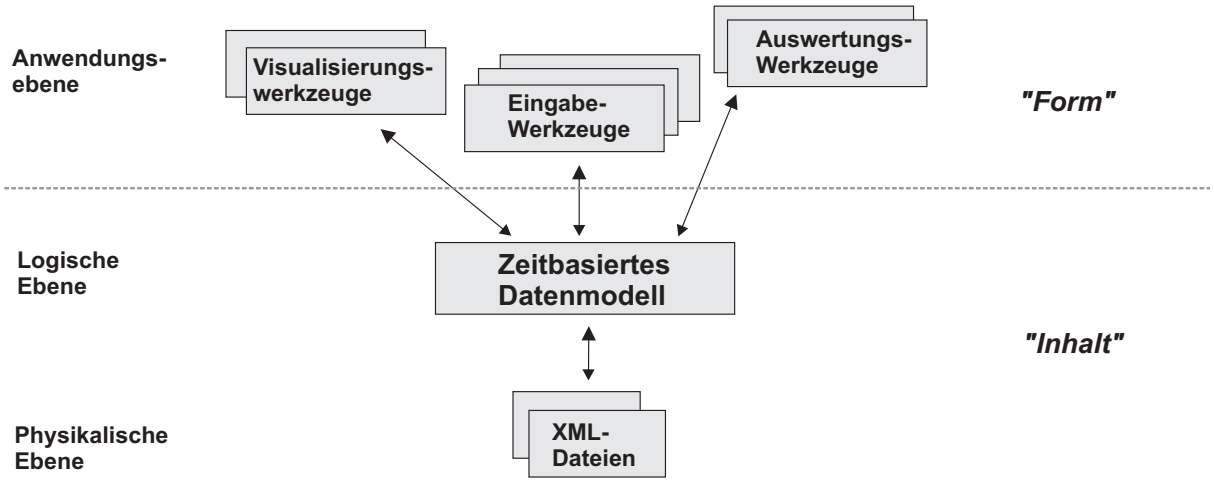
3.1 Prinzipien

Jenseits der Details der konkreten Implementierung haben sich im Laufe der nun vierjährigen Projektarbeit einige grundlegenden Prinzipien herauskristallisiert, die als Leitlinien bei der Entwicklung der Datenbank Mehrsprachigkeit dienen

und von denen wir glauben, dass sie auch in anderen Zusammenhängen, in denen es um die computergestützte Verarbeitung heterogener linguistischer Daten geht, von Nutzen sein mögen. Diese Prinzipien sind im Folgenden zusammengefasst.

3.1.1 Drei-Ebenen-Architektur

Die Datenbank Mehrsprachigkeit geht von einer Drei-Ebenen-Architektur der computergestützten Datenverarbeitung aus:



Diese beinhaltet zunächst eine *Trennung von Form und Inhalt* von Dokumenten, die im Rahmen texttechnologischer Verfahren mittlerweile als selbstverständlich gilt, innerhalb der methodologischen Grundlagen linguistischer Transkription aber bislang noch weitestgehend unbeachtet geblieben ist. Die Form von Transkriptionen betrifft ihre graphische Darstellung, z.B. für eine gedruckte Ausgabe auf Papier oder für die Anzeige in einem Transkriptions- oder Auswertungswerkzeug auf dem Computerbildschirm. Viele Gegensätze bestehender Transkriptionsverfahren, beispielsweise die jeweils favorisierte Notationsform (z.B. Partitur- vs. Zeilennotation) oder die Darstellung spezifischer Gesprächsphänomene (z.B. „Pausenzeichen“), sind allein der Formebene zuzurechnen. Der Inhalt von Transkriptionen besteht hingegen aus einer Menge von symbolisch beschriebenen Gesprächseinheiten sowie deren Beziehungen zueinander. Relevante Unterschiede zwischen Transkriptionssystemen auf der Inhaltsebene be-

treffen beispielsweise die Definition und Benennung von Gesprächseinheiten (z.B. Äußerungen vs. Phrasierungseinheiten) und den Detailliertheitsgrad der Markierung von zeitlichen Relationen (s.o.). Durch eine konsequente Trennung von Form und Inhalt kann bereits eine wesentlich erhöhte Flexibilität im Umgang mit Transkriptionsdaten erzielt werden, denn sie ermöglicht es, ein und dasselbe Dokument für unterschiedliche Zwecke auf unterschiedliche Weise zu visualisieren. Darüber hinaus bietet sie die Möglichkeit, rein formbasierte Differenzen zwischen verschiedenen Transkriptionssystemen durch eine einheitliche Repräsentation auf der Inhaltsebene aufzuheben.

Über die Trennung von Form („Anwendungsebene") und Inhalt („Datenebene") hinaus sieht die Drei-Ebenen-Architektur eine weitere *Unterscheidung zwischen logischer und physikalischer Datenstruktur* vor. Die logische Struktur beschreibt unabhängig von konkreten technologischen Umgebungen die grundlegenden Organisationsprinzipien für Transkriptionsdaten in Form eines Datenmodells. Beispielsweise wird im Annotationsgraphen-Formalismus (Bird/Liberman 2001) vorgeschlagen, Transkriptionsdaten auf der logischen Ebene als azyklische gerichtete Graphen zu beschreiben, während das NITE-Object-Model (Evert et al. 2003) von einem System überlappender Hierarchien als grundlegender struktureller Organisationsform ausgeht. Auf der physikalischen Ebene hingegen wird festgelegt, wie diese abstrakten Datenstrukturen als computerlesbare Dateien zu kodieren sind. Wie Bird/Liberman (2001) feststellen, ist nur durch eine Trennung von logischer und physikalischer Datenebene sicherzustellen, dass Transkriptionsdaten über spezifische technologische Umgebungen hinaus langfristig nutzbar und austauschbar bleiben.[1]

[1] Gegenwärtig überdeckt der flächendeckende Einsatz von XML diesen Umstand häufig. Da XML oft nicht nur als Standard für die physikalische Repräsentation strukturierter Daten angesehen wird, sondern eng mit einem zugehörigen logischen (OHCO-)Datenmodell assoziiert ist, vernachlässigen einige aktuelle Ansätze diese essentielle Unterscheidung. Ge-

3.1.2 Aspekte der „Standardisierung“

Im Zusammenhang mit der computergestützten Verarbeitung heterogener Datenbestände wird oft deren „Standardisierung“ als grundlegendes Desiderat genannt. Die obigen Ausführungen zu konzeptuell vs. technologisch bedingter Heterogenität und zur Drei-Ebenen-Architektur der Datenverarbeitung können helfen, verschiedene Aspekte dieses Begriffs zu differenzieren:

Für diejenigen Unterschiede zwischen Daten, die sich aus konzeptuellen Überlegungen motivieren, verbietet sich eine Standardisierung. Weil unterschiedliche Forschungsziele und theoretische Hintergründe teilweise unterschiedliche Datenformen zwingend erfordern, kann das Ziel einer projektübergreifend einsetzbaren Lösung nicht sein, eine vollständig vereinheitlichte Form für Transkriptionsdaten vorzuschlagen. Vielmehr muss sich eine solche Lösung darauf beschränken, auf der Basis struktureller Gemeinsamkeiten verschiedener Systeme ein abstraktes „Framework“ zu erarbeiten, das möglichst wenige ontologische Festlegungen[2] trifft und für verschiedene theoretische Herangehensweisen parametrisierbar ist.

Während auf der logischen Ebene der Datenverarbeitung eine Standardisierung also auf ein solch abstraktes Framework begrenzt bleiben muss, bieten sich auf der physikalische Ebene weit reichende Möglichkeiten für die Nutzung von Standards: viele praktische Probleme in der Arbeit mit heterogenen Datenbeständen ergeben sich weniger aus deren prinzipieller konzeptueller Inkompatibilität, sondern vielmehr aus der Tatsache, dass sie in proprietären (binären oder textbasierten) und damit schwer zu verarbeitenden Formaten vorliegen. Der

---

rade für Transkriptionsdaten, deren grundlegende strukturelle Merkmale (insb. parallele Strukturen) sich nicht vollständig in das „Standard-XML-Datenmodell“ einordnen, ist es m.E. jedoch wichtig anzuerkennen, dass mit der Wahl von XML als Speicherformat noch keine Entscheidung über eine logische Datenstruktur getroffen ist.

[2] Auch Bird/Liberman (2001) bezeichnen ihren Annotationsgraphen-Ansatz als „ontologically parsimonious“.

einheitliche Einsatz von XML und Unicode kann daher auf dieser Ebene bereits zu einer entscheidenden Erleichterung der Verarbeitung beitragen.[3]

Auf der Anwendungsebene hingegen scheint eine Standardisierung weder wünschenswert noch notwendig. Eine Vielfalt von Darstellungsmöglichkeiten für Transkriptionsdaten kommt der Forschungspraxis ebenso entgegen wie die Möglichkeit, ein und dasselbe Datum mit unterschiedlichen Software-Werkzeugen bearbeiten zu können, und die Trennung von Form und Inhalt von Dokumenten stellt sicher, dass diese Vielfalt auf der Anwendungsebene keine Inkompatibilitäten auf der Datenebene nach sich ziehen muss.

3.1.3 Berücksichtigung bewährter Arbeitsweisen

Die Konstruktion der Datenbank Mehrsprachigkeit findet in einem Umfeld statt, in dem sich bereits viele verschiedene Ansätze für die computergestützte Verarbeitung von Transkriptionen gesprochener Sprache – teilweise über viele Jahre hinweg – etabliert haben. Zwar verspricht das Projektziel einen offensichtlichen qualitativen Sprung gegenüber all diesen Ansätzen; dennoch ist die Akzeptanz der entwickelten Lösungen in hohem Maße davon abhängig, dass bewährte Arbeitsweisen nicht übergangslos über Bord geworfen werden.

So müssen bereits beim Entwurf von Datenmodellen und -formaten Zugeständnisse an Unzulänglichkeiten älterer Datenbestände gemacht werden, denn deren Überführbarkeit stellt eine unabdingbare Voraussetzung für das Erreichen der Projektziele dar. Die Konvertierung von „Legacy Data" ist daher alles andere als eine triviale, rein technologische Aufgabe – sie führt notgedrungen zu ei-

---

[3] Tatsächlich ist nach unserer Erfahrung der Schritt von einem beliebigen textbasierten oder binären Format zu einem XML-basierten Format in der Regel um ein Vielfaches aufwändiger als eine Überführung eines XML-Datums in ein anderes XML-Datum. Da XML sich flächendeckend durchzusetzen scheint, steht zu hoffen, dass viele der momentan akuten Probleme im Umgang mit digitalen Sprachressourcen in Zukunft obsolet sein werden.

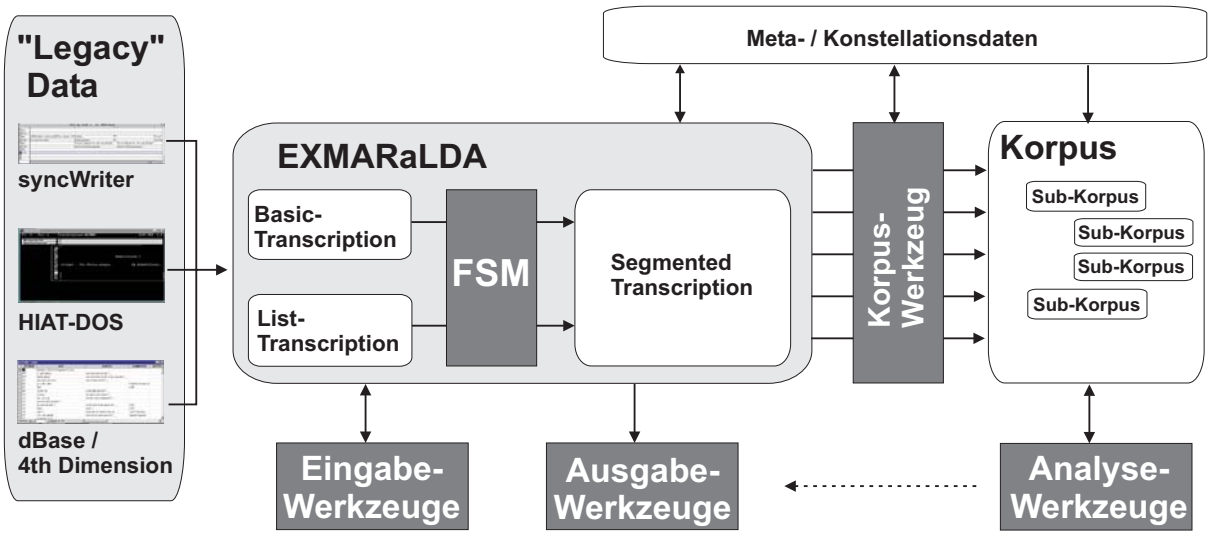
ner Reihe von Kompromissen und bestimmt so in entscheidender Weise die Architektur des zu entwickelnden Systems mit.

Unter den Aspekt der Beibehaltung bewährter Arbeitsweisen fällt auch die Berücksichtigung solcher Analyseschritte, die nur am Rande einer *computergestützten* Korpusarbeit zuzurechnen sind: so beinhalten diskursanalytische Verfahren als einen wichtigen methodischen Schritt eine intensive qualitative Analyse eines *gedruckten* Transkripts. Die üblicherweise bildschirmzentrierten Verfahren computergestützter Korpusarbeit müssen daher um die (insbesondere im Falle der Partiturnotation technisch anspruchsvolle) Möglichkeit der Ausgabe gedruckter Visualisierungen ergänzt werden.Schließlich hat die Orientierung an bewährten Arbeitsweisen auch dazu geführt, dass für die Datenbank Mehrsprachigkeit generell ein „Bottom-Up"-Konzept verfolgt wird, das eine verteilte Erstellung von einzelnen Transkriptionen und Korpora einer „zentralistischen" Datenverwaltung vorzieht. Die in vielen vergleichbaren Projekten angestrebte „Top-Down"-Lösung, nach der die Zusammenführung verschiedener Datenbestände in einer gemeinsamen Oberfläche ein übergeordnetes Ziel darstellt, hat sich für die Forschungspraxis am SFB 538 als nicht praktikabel erwiesen. Ausschlaggebend dafür waren einerseits Vorbehalte der einzelnen Forscher, die teilweise ihre Kontrolle über Persönlichkeits- und Urheberrechte bzgl. der Daten gefährdet sahen. Andererseits ließ auch die Prämisse, entwickelte Werkzeuge möglichst einfach Personen außerhalb des SFB zur Forschung und Lehre zur Verfügung stellen zu können, eine Lösung sinnvoll erscheinen, in der einzelne Bestandteile des Systems möglichst unabhängig von einer übergeordneten Architektur nutzbar sind.[4]

[4] Natürlich schließt eine solche „Bottom-Up"-Lösung nicht aus, dass dezentral erstellte Daten in übergeordneten Strukturen zusammengefasst und zugänglich gemacht werden. Sie beschränkt sich aber darauf, die *Voraussetzungen* für einen solchen Schritt zu schaffen und

3.2 Systemarchitektur

Folgende Abbildung illustriert die Systemarchitektur der Datenbank Mehrsprachigkeit:



Auf eine detaillierte Darstellung der Einzelkomponenten soll an dieser Stelle verzichtet werden. Statt dessen mögen die folgenden Ausführungen deutlich machen, wie diese Architektur mit den im vorigen Abschnitt angeführten Prinzipien zusammenhängt:

Die zentrale Komponente der Systemarchitektur ist das EXMARaLDA-Datenmodell. Dieses stellt eine eingeschränkte und spezifizierte Version des Bird/Libermannschen Annotationsgraphen-Datenmodells dar, geht also davon aus, dass sich Transkriptionsdaten angemessen als azyklische gerichtete Graphen auffassen lassen, deren Knoten den gemeinsamen Zeitbezug aller transkribierten Einheiten repräsentieren und deren Kanten die nicht-zeitlichen Informationen tragen. Diesem zeitbasierten Datenmodell auf der logischen Ebene entsprechen auf der physikalischen Ebene drei aufeinander aufbauende XML-

belässt die Entscheidung und Kontrolle über die konkrete Form einer übergeordneten Datenverwaltung bei den Einzelprojekten.

Formate: Die volle strukturelle Komplexität des Datenmodells kann in einer *Segmented-Transcription* repräsentiert werden. *Basic-Transcription* und *List-Transcription* bilden jeweils strukturell vereinfachte Untermengen einer solchen *Segmented-Transcription*. Die primären Eingabeinstrumente (insb. der Partitur-Editor) können der Effizienz halber zunächst auf diesen vereinfachten Untermengen operieren und nach Abschluss des Eingabeprozesses die Daten automatisch in das mächtigere *Segmented-Transcription*-Format überführen.[5] Die Überführung vorhandener Datenbestände besteht zunächst in der Abbildung von deren Strukturen auf das zeitbasierte Datenmodell und dann in der konkreten Konvertierung der jeweiligen Dateien in die entsprechenden XML-Formate.

Datenmodell und -formate sind gemäß den obigen Überlegungen prinzipiell unabhängig von Präsentationsformaten und Bearbeitungssoftware. Geeignete Ein- und Ausgabewerkzeuge werden teilweise im Projekt selbst entwickelt (EXMARaLDA Partitur-Editor), die Systemarchitektur sieht aber ausdrücklich vor, dass auch andernorts entwickelte Software, die auf ähnlichen Datenmodellen operiert (Praat, TASX-Annotator und ELAN), für die Erstellung von EXMARaLDA-Daten verwendet werden kann.

Zusammen mit solchen Ein- und Ausgabewerkzeugen bildet EXMARaLDA bereits ein selbstständig, d.h. ohne weitere übergeordnete Komponenten, nutzbares System und wird als solches auch vielfach in Forschung und Lehre eingesetzt. Um dem Ziel gerecht zu werden, eine gemeinsame Plattform für die verschiedenen SFB-Projekte zu bilden, kann es jedoch durch weitere Komponenten ergänzt werden: Ein Korpus-Werkzeug (CoMa, EXMARaLDA Corpus-Manager) erlaubt die Bündelung mehrerer EXMARaLDA-Transkriptionen zu

[5] Der in der Praxis hierfür benutzte Mechanismus ist eine Finite State Machine (FSM), die sich die Regelmäßigkeiten der verwendeten Transkriptionssysteme zunutze macht, um in den transkribierten Symbolketten implizite Markierungen in explizite Strukturrepräsentationen umzuwandeln.

einem Korpus, das wiederum in Form einer XML-Datei physikalisch repräsentiert wird. Ein in der Entwicklung befindliches Analysewerkzeug (SQUIRREL, Search and Query Instrument for EXMARaLDA) operiert dann auf Untermengen solcher Korpora, die anhand einer Suche auf Meta- und Konstellationsdaten (z.B. Sprechereigenschaften, verwendete Sprachen, Diskurstyp) ausgewählt werden.

4 Praktische Erfahrungen

Obwohl ein erheblicher Teil der sprachwissenschaftlichen Methoden sich heutzutage auf eine computergestützte Verarbeitung von Sprachkorpora stützt, und obwohl innerhalb der vergangenen fünfzehn Jahre eine Vielzahl entsprechender Werkzeuge und Datenformate entstanden ist, bleibt die Entwicklung von computergestützten Systemen für die linguistische Forschung ein Thema, das bislang kaum Eingang in die wissenschaftliche Literatur gefunden hat. Bei Beginn des Projekts „Datenbank Mehrsprachigkeit" bestanden demzufolge lediglich vage Vorstellungen über den zeitlichen und personellen Aufwand und die potentiellen inhaltlichen und technologischen Schwierigkeiten, die ein solches Vorhaben mit sich bringt. Im Laufe der nunmehr vierjährigen Projektarbeit haben sich diese Vorstellungen – in einem teilweise mühsamen Lernprozess – konkretisiert, und das Folgende ist der Versuch, einige der wesentlichen diesbezüglichen Erfahrungen festzuhalten.

4.1 Entwicklungsarbeit

4.1.1 Technologien

Hinsichtlich der zu verwendenden Technologien wurden bereits zu Projektbeginn drei verbindliche Entscheidungen getroffen:

Grundlage der physikalischen Datenrepräsentation sollten XML und Unicode sein, weil beide als offene Standards und aufgrund ihrer sich anbahnenden

Akzeptanz in der gesamten Internet-Welt einen Ausweg aus dem Problem der mangelnden Archivierbarkeit von Transkriptionsdaten versprachen. Dieses Versprechen ist eingehalten worden. XML- und Unicode-Technologie wird mittlerweile zuverlässig von einer Vielzahl von Werkzeugen und Programmierbibliotheken unterstützt, und der weitaus größte Teil vergleichbarer Projekte weltweit sieht ebenfalls XML- und Unicode-basierte Lösungen für die physikalische Repräsentation von digitalen Sprachdaten vor. Gegenüber der Ausgangssituation, in der die Vielfalt an proprietären (und teilweise kaum dokumentierten) Formaten und Kodierungen einfachste Verarbeitungsschritte oft unmöglich machte, stellt dies einen kaum zu überschätzenden Fortschritt dar. Weitere wesentliche Verbesserungen wären aus unserer Sicht vor allem dann zu erwarten, wenn auf XML aufbauende Technologien (insbesondere XSLT, XSL:FO, SMIL) einen der XML-Kerntechnologie vergleichbaren Grad der Unterstützung und Zuverlässigkeit erreichen würden.

Als Grundlage für die Implementierung der Software wurde JAVA ausgewählt. Damit verband sich vor allem die Erwartung, mit vertretbarem Aufwand Werkzeuge entwickeln zu können, die plattformübergreifend – insbesondere in der Windows- *und* der Macintosh-Welt – einsetzbar sind. Grundsätzlich ist auch diese Erwartung erfüllt worden – alle EXMARaLDA-Werkzeuge sind auf verschiedenen Betriebssystemen lauffähig –, allerdings hat sich die JAVA-Philosophie des „Write once, run anywhere“ stellenweise nicht bewahrheitet. Dies liegt einerseits darin begründet, dass die Macintosh-Implementierung der Java-Maschine bis heute unter gelegentlicher Instabilität und mangelnder Dokumentation leidet, was entsprechende fortwährende betriebssystemspezifische Anpassungen des Codes notwendig macht. Andererseits scheint besonders die im Zusammenhang mit der Transkription gesprochener Sprache wichtige Arbeit mit digitalisierten Medien-Signalen (Audio und Video) ein Teilbereich zu sein, der von „hardware-fernen“ Technologien wie JAVA prinzipiell nicht optimal

unterstützt werden kann.[6] Auf lange Sicht wünschenswert wären in dieser Hinsicht plattformspezifische Lösungen mit entsprechenden Interfaces zu JAVA.

4.1.2 Entwicklungsphasen / Zeitlicher Aufwand

Eine weitestgehend unbekannte Größe zu Projektbeginn war der für die einzelnen Phasen der Entwicklungsarbeit zu veranschlagende zeitliche Aufwand. Unterscheidet man nach der gängigen Praxis des Software-Engineering in etwa die folgenden Phasen der Entwicklungsarbeit – Planung/Analyse/Entwurf, Implementierung, Test, Dokumentation, Wartung, Benutzersupport –, so hat die Projektarbeit deutlich gezeigt, dass der zeitliche Aufwand für die „eigentliche" Programmierung (d.h. Implementierung und Wartung) der Software von den übrigen Größen um ein Vielfaches übertroffen wird.

Zu Projektbeginn gestaltete sich zunächst die Definition eines Anforderungsprofils für die zu entwickelnden Systemkomponenten sehr aufwändig. Die Erwartungen der potentiellen Benutzer beschränkten sich zunächst auf sehr allgemeine Anforderungen (wie „Benutzerfreundlichkeit der Software", „Hilfe bei quantitativen Analysen") und konnten auch in mehreren „Brainstorming"-Treffen nicht hinreichend spezifiziert werden. Es wurde daher zunächst auf der Basis einer vorläufigen Liste von wünschenswerten Merkmalen ein Prototyp eines Transkriptionseditors implementiert und interessierten Personen zur Verfügung gestellt. Dabei (und auch im folgenden Projektverlauf) zeigte sich, dass das Testen und kritische Begutachten von Beta-Software eine Tätigkeit ist, die einer Aufmerksamkeit und Sorgfalt bedarf, für die in der alltäglichen Forschungspraxis kaum Raum zu sein scheint – es erwies sich als sehr schwierig und zeitaufwändig, Forscher zur Auseinandersetzung mit einer Software zu be-

---

[6] Technologien wie das „Java Media Framework" und „Java Sound API" bieten zwar eine durchaus brauchbare Unterstützung für grundlegende Funktionen in dieser Hinsicht. Sie arbeiten nach unserem Eindruck jedoch merklich weniger präzise und zuverlässig als Lösungen, die in Sprachen wie C++ o.ä. implementiert wurden.

wegen, die aufgrund ihres frühen Entwicklungsstadiums keinen unmittelbaren praktischen Nutzen für die aktuell anstehende Forschungsarbeit zu versprechen vermochte. In diesem Sinne hat sich die ursprüngliche Erwartung, dass der allseits geäußerte dringende Bedarf an einer zeitgemäßen Transkriptions-Software von sich aus zu einer Vielzahl von Testern führen würde, als trügerisch erwiesen. Entscheidende Abhilfe schuf hier erst die Einstellung von Hilfskräften, die explizit mit dem Abfassen von Test-Berichten beauftragt wurden.

Nach diesen anfänglichen Schwierigkeiten hat sich inzwischen ein zirka drei- bis viermonatiger Zyklus etabliert, in dem neue Software-Versionen über die Projekt-Website veröffentlicht, anschließend Rückmeldungen über Bugs und Verbesserungsvorschläge gesammelt und diese in die Software eingearbeitet werden. Zu beobachten ist dabei, dass grundlegende Funktionserweiterungen wesentlich langsamer wahrgenommen werden als Umgestaltungen in bereits vorhandenen Komponenten. Die Zahl der „Power-User“, also von Personen, die neue Programmfunktionen in ihrem vollem Umfang frühzeitig und intensiv nutzen, ist vergleichsweise gering; der Großteil der Benutzer zeigt sich eher an der Optimierung von Vorhandenem interessiert. Der wichtigste zeitliche Faktor bei der Weiterentwicklung der Software ist aber mittlerweile weniger die Definition und Implementierung der Änderungen und Erweiterungen selbst, sondern deren Dokumentation in Form von Benutzerhandbüchern und Beispielen. Ähnliches gilt für die individuelle Beratung von Nutzern (vornehmlich über E-Mail), die einerseits zwar häufig nur die in der schriftlichen Dokumentation enthaltene Information dupliziert, anderseits aber auch entscheidend dazu beigetragen hat, dass mittlerweile eine wesentlich konkretere Vorstellung über den tatsächlichen und potentiellen Nutzerkreis der Software besteht.

4.2 Benutzer

Obwohl das Kernziel des Projektes in der Entwicklung einer Lösung für die Projektarbeit *am SFB 538* besteht, hat die EXMARaLDA-Software (insb. der Partitur-Editor) inzwischen eine recht weite Verbreitung über den SFB hinaus gefunden. Da bis vor kurzem der Download eine schriftliche Anmeldung voraussetzte (und über die Erfahrungen aus dem individuellen Benutzersupport, s.o), besteht zumindest eine ungefähre Vorstellung darüber, wie sich der derzeitige EXMARaLDA-Benutzerkreis zusammensetzt: Nach einer vorsichtigen Schätzung wurden seit Dezember 2001 (Version 1.0. des Editors) ca. 800 Benutzerkennungen angefordert. Weit über die Hälfte davon stammten von Studierenden, die EXMARaLDA im Rahmen einer sprachwissenschaftlichen Lehrveranstaltung nutzten, dies zum allergrößten Teil an deutschen Universitäten, teilweise aber auch im Ausland, vor allem in der Schweiz und den USA. Ebenfalls in der Lehre kommt EXMARaLDA bei der Lehramtsausbildung für Mathematiker und in der Kommunikationsforschung zum Einsatz. Projekte, die EXMARaLDA in der Forschung einsetzen, verfolgen zum überwiegenden Teil gesprächsanalytische Fragestellungen, weitere Anwendungsfelder finden sich in der Spracherwerbsforschung und in handlungsanalytisch orientierten erziehungswissenschaftlichen Projekten.

Wie bereits erwähnt, hat sich die aus dieser weiten Verbreitung resultierende hohe Zahl von kritischen Rückmeldungen bereits positiv auf die Entwicklungsarbeit ausgewirkt. Darüber hinaus erlaubt sie erste Mutmaßungen darüber, wie die Entwicklung computergestützter Systeme in der derzeitigen Forschungslandschaft aufgenommen wird. Zwei Aspekte scheinen mir hierbei besonders

wichtig, und das Folgende ist ein Versuch, diese in der gebotenen Kürze (und Vorsicht[7]) zu formulieren:

4.2.1 Die Rolle des Computers in der linguistischen Methode

Man kann den Einsatz des Computers für sprachwissenschaftliche Untersuchungen unter zwei Gesichtspunkten betrachten: zum einen kann der Computer als ein technisches Instrument gesehen werden, das vornehmlich dazu dient, gewisse Arbeitsschritte, die prinzipiell auch ohne seine Hilfe durchführbar wären, zu vereinfachen.[8] Zum anderen können computergestützte Verfahren aber auch als eine grundsätzliche Erweiterung des wissenschaftlichen Methodenrepertoires aufgefasst werden, etwa indem der Rechner als ein Instrument zum Anfertigen und Manipulieren wissenschaftlicher Modelle gesprochener Sprache betrachtet wird.[9] Nach unserer Erfahrung ist im Bereich der Gesprächs- und Spracherwerbsforschung, in denen EXMARaLDA vornehmlich zum Einsatz kommt, die erste Sicht die eindeutig vorherrschende. Der Nutzen neuer Lösungen wird weniger danach beurteilt, welche neuen Möglichkeiten sie bieten, sondern eher danach, wie sie bestehende Methoden zu unterstützen vermögen. Konkret äußert sich dies in der bereits angesprochenen Zurückhaltung der Nutzer beim Erpro-

---

[7] Die hierbei unvermeidlichen und eigentlich unzulässigen Verallgemeinerungen bitte ich nachzusehen. Dass sich diese subjektiven Eindrücke kaum „objektiv“ durch Verweise auf eine öffentliche wissenschaftliche Diskussion belegen lassen, ist Teil des Problems, das hier thematisiert werden soll.

[8] Beim Transkribieren betreffen solche Vereinfachungen z.B. die (auf dem Papier aufwändigeren) iterativen Korrekturschritte, das (auf dem Papier teure und u.U. nicht verlustfreie) Vervielfältigen und Verteilen von Transkripten, das (bei analogen Geräten oft umständliche und verschleißbehaftete) Abspielen der zu transkribierenden Aufnahme oder das (ohne Computerunterstützung mühselige und u.U. unzuverlässige) Suchen nach sprachlichen Phänomenen in größeren Korpora.

[9] Orlandi (2002) bringt diese unterschiedlichen Auffassungen wie folgt zum Ausdruck: “[Some] colleagues refer to the computer as ‘just a tool’ or ‘simply a bunch of techniques’, as if ways of knowing did not have much to do with what is known. Because the computer is a meta-instrument – a means of constructing virtual instruments or models of knowing – we need to understand the effects of modelling on the work we do as humanists.” Vgl. dazu auch Schmidt (i.V.)

ben solcher Funktionalitäten, die nicht bereits aus vorhandenen Systemen bekannt sind, aber auch darin, dass – bis auf wenige Ausnahmen – die Rolle computergestützter Verfahren in den methodologischen Grundlagen der genannten Gebiete bislang weitestgehend unreflektiert bleibt.[10]

4.2.2 Kooperative Korpusarbeit

Ein leitender Gedanke bei der Konstruktion der Datenbank Mehrsprachigkeit ist die Überwindung von Hindernissen, die einem projektübergreifenden Zugriff auf Korpora gesprochener Sprache derzeit im Wege stehen. Dies erscheint einerseits aus rein ökonomischen Gesichtspunkten wünschenswert, denn die Erstellung von Aufnahmen und Transkriptionen ist bekanntermaßen mit hohem finanziellem und personellem Aufwand verbunden, der sich umso eher rechtfertigen lässt, je mehr Forschende auf die solchermaßen entstandenen Ressourcen zugreifen können. Andererseits sprechen auch gegenstandstheoretische Gründe dafür, Korpusarbeit als eine kooperative Aufgabe wahrzunehmen, denn oft kann nur durch die Zusammenlegung verschiedener Korpora die „kritische Masse" erzielt werden, die für eine aussagekräftige (u.U. statistisch untermauerte) quantitative Analyse sprachlicher Phänomene notwendig ist.

Im Bereich der Sprachtechnologie hat diese Erkenntnis bereits zu einer ganzen Reihe von Initiativen und Institutionen geführt, die sich der Bereitstellung einer organisatorischen und technischen Infrastruktur für den Austausch von digitalen Sprachressourcen widmen.[11] Innerhalb der nicht technologisch ausgerichteten Linguistik wird die Zweckmäßigkeit solcher Bemühungen zwar nicht grundsätzlich in Frage gestellt; es hat bis heute aber weder eine nennens-

---

[10] Hingegen mangelt es nicht an Reflexionen über die Methode der Transkription als solcher. Gerade auf diesem Gebiet versprechen moderne texttechnologische Methoden – z.B. die o.g. Trennung von Inhalt und Form von Transkriptionen – aber eine grundlegende Erweiterung der Möglichkeiten etablierter Verfahren.

[11] Z.B. Organisationen wie ELDA oder LDC und Projekte wie EAGLES/ISLE oder ATLAS.

werte Anbindung an solche Initiativen stattgefunden, noch existieren eigenständige Konzepte, um die mancherorts entwickelten Einzellösungen (zu denen z.B. die CHILDES-Datenbank zählt) in praktikabler Weise miteinander zu verbinden. Projekte wie das hier vorgestellte (und weitere der auf dem Workshop präsentierten Arbeiten) können zwar gewisse Voraussetzungen für eine solche Infrastruktur schaffen, indem sie zumindest innerhalb der Institutionen, an denen sie angesiedelt sind, einen gemeinsamen Überbau für die Korpusarbeit entwerfen. Mindestens ebenso wichtig wäre jedoch eine dezidierte Bereitschaft der beteiligen Forscher, Fragen der Austauschbarkeit und Archivierbarkeit von Sprachdaten von vorneherein in die Korpuserstellung einzubeziehen, und die Entwicklung von Infrastrukturen, innerhalb derer Korpora anderen Interessierten zur Verfügung gestellt werden können, aktiv zu unterstützen. Die in Bird/Simons (2002) ausgeführte Beobachtung, dass diese Bereitschaft nicht uneingeschränkt gegeben ist, weil viele Forscher die Nachteile kooperativer Korpusarbeit höher bewerten als die sich aus ihr ergebenden Vorteile[12], können wir bestätigen.

5 Zusammenfassung und Ausblick

Wie andere vergleichbare Arbeiten zeigt auch das Beispiel der Datenbank Mehrsprachigkeit und EXMARaLDA, dass die Anwendung texttechnologischer Methoden und Konzepte und der Einsatz standardisierter und plattformübergreifend nutzbarer Technologien einen wesentlichen Fortschritt für die Arbeit mit heterogenen linguistischen Daten mit sich bringen kann. Das vornehmliche Ziel des vorliegenden Aufsatzes ist jedoch darauf hinzuweisen, dass dadurch nach unse-

[12] Dies betrifft ganz besonders die Frage der Zitierfähigkeit wissenschaftlicher Primärdaten. Bird/Simons (2002) sagen dazu: „Commonly, a researcher wants to derive recognition for the labor that went into creating primary language documentation, but does not want to make the materials available to others until deriving maximum personal benefit."

rer Erfahrung mindestens ebenso viele neue Fragen aufgeworfen wie alte beantwortet werden. Die nunmehr vierjährige Projektarbeit hat nämlich immer deutlicher werden lassen, dass technologische Innovation (sprich: Softwarewerkzeuge, Datenmodelle und -formate) nur eines von drei Standbeinen ist, auf die sich eine solcher Fortschritt stützt. Weitere entscheidende Verbesserungen sind zu erwarten, wenn die Möglichkeiten, die durch technische Weiterentwicklungen eröffnet werden, von einer entsprechenden methodologischen Reflexion begleitet und durch den Aufbau einer geeigneten Infrastruktur innerhalb der betroffenen Forschergemeinden unterstützt werden. Idealerweise würde dazu die Arbeitsteilung zwischen Texttechnologen und Informatikern, die Software und Datenformate entwickeln, und Sprachwissenschaftlern, die diese anwenden, teilweise aufgehoben oder zumindest stärker als bisher durch einen interdisziplinären Dialog ergänzt.

6 Literatur

Bird, Steven / Liberman, Mark (2001): *A formal framework for linguistic annotation.* In: Speech Communication 33(1,2), 23-60.

Bird, Steven / Simons, Gary (2002): *Seven Dimensions of Portability for Language Documentation and Description.* In: Language 79, 557-582.

Evert, Stefan / Carletta, Jean / O'Donnell, Timothy J. / Kilgour, Jonathan / Vögele, Andreas / Voormann, Holger (2003): The NITE Object Model. Version 2.1. (24 March 2003). NITE Internal document. http://www.ltg.ed.ac.uk/NITE/documents.html

Orlandi, Tito (2002): Is humanities computing a discipline? In: Braungart, Georg / Eibl, Karl / Jannidis, Fotis (Hrsg.) (2002): Jahrbuch für Computerphilologie 4. Paderborn: Mentis, 51-58.

Rehbein, Jochen / Schmidt, Thomas / Meyer, Bernd / Watzke, Franziska / Herkenrath, Annette (2004): Handbuch für das computergestützte Transkribieren nach HIAT. Arbeiten zur Mehrsprachigkeit, Serie B (56). Hamburg.

Schmidt, Thomas (2004): Transcribing and annotating spoken language with EXMARaLDA. In: Witt, Andreas / Heid, Ulrich / Carletta, Jean / Thompson, Henry S. / Wittenburg, Peter (Hrsg.): XML-based richly annotated corpora. LREC 2004 Satellite Workshop. Paris: ELRA, 69-74.

Schmidt, Thomas (i.V.): Computergestützte Transkription als Modellierung und Visualisierung gesprochener Sprache mit texttechnologischen Mitteln. Dissertation, Universität Dortmund.

Thomas Schmidt
Universität Hamburg
SFB 538 „Mehrsprachigkeit“
Max Brauer-Allee 60
22765 Hamburg
Germany
thomas.schmidt@uni-hamburg.de
http://www.rrz.uni-hamburg.de/exmaralda